High Bandwidth Memory (HBM2) IP Subsystem Silicon Validation And Interoperability With HBM2 Memory Die Stack

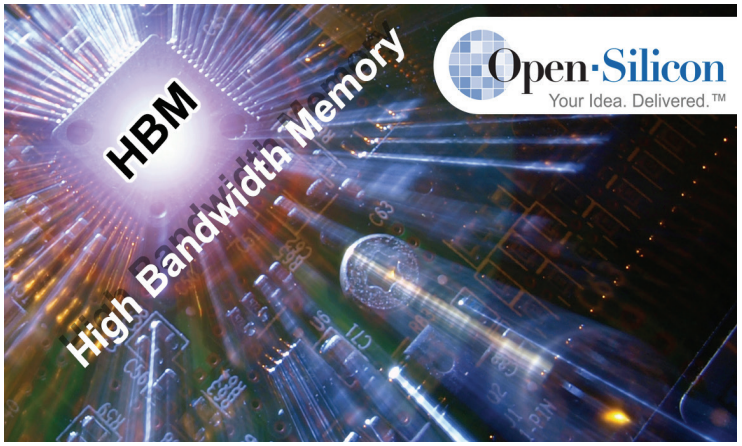
Open-Silicon

ABSTRACT

The most common memory requirements for emerging applications, such as high performance computing, networking, deep learning, virtual reality, gaming, cloud computing and data centers, are high bandwidth and density, based on real-time random operations. High Bandwidth Memory (HBM2) meets these requirements and delivers unprecedented bandwidth, power efficiency and small form factor. HBM2 (x1024) offers the maximum possible bandwidth of up to 256GBps, compared to 4GBps with DDR3 (x16), at 1/3rd of the power efficiency. One of the key IPs used to develop these ASIC system in packages (SiPs) is the HBM IP subsystem that consists of a controller, PHY and die-to-die I/O. Open-Silicon's HBM2 IP subsystem fully complies with the HBM2 JEDEC® standard. The IP translates user requests into HBM command sequences (ACT, Pre-Charge) and handles memory refresh, bank/page management and power management on the interface. The high performance, low latency controller leverages the HBM parallel architecture and protocol efficiency to achieve maximum bandwidth. The IP includes a scalable and optimized PHY and die-to-die custom I/O needed to drive the interface between the logic die and the memory die stack on the 2.5D silicon interposer. Open-Silicon's HBM2 IP subsystem addresses the implementation challenges associated with interoperability, 2.5D design, overall SiP design, packaging, test and manufacturing. Multiple built-in test and diagnostic features, such as probe pads and loop-back for issue-isolation within the various IP subsystem components, not only address the test and debug challenges, but help in yield management and yield improvement, all while ramping HBM2 ASIC designs into volume production. Open-Silicon's first HBM2 implementation solution in TSMC's 16nm FF+ features 2Gbps per pin data rate at up to 5mm trace length. This enables a full 8-channel connection from a 16nm SoC to a single HBM2 memory stack at 2Gbps, achieving bandwidths up to 256GBps.

During this presentation, Open-Silicon will present silicon validation results of a 2.5D HBM2 ASIC SiP test platform, which is based on Open-Silicon's HBM2 IP subsystem in TSMC's 16nm FF+ process in combination with TSMC's CoWoSTM 2.5D silicon- interposer technology and Samsung's HBM2 memory. The test platform has a built-in data generator engine to drive the user interface (referred as "UIF") of the HBM2 controller. This data generator engine can generate either a fixed or variable data pattern using a pseudo random binary sequence (PRBS). A built-in data monitor on the UIF is used to compare the expected data. It utilizes four UIF interfaces to create traffic on four pseudo channels of the HBM2 memory. It executes random/increment/decrement/walking zero/walking one addressing patterns. These different addressing patterns cover multiple traffic patterns used by high bandwidth applications. For example, an incremental addressing pattern is evident in video buffer traffic, a random addressing pattern is evident in network packet traffic, and so on. After execution, the test platform will capture test status statistics, bandwidth and latency (min, max and average). For functional validation and IP subsystem-to-HBM2 memory die stack interoperability, results from the data generated from the four different combinations of back-to-back transactions on all four UIFs are displayed together on the Graphical User Interface (GUI) running on the PC. Silicon validation provides results that include a calibration report on a valid data window validating signal integrity, power consumption of HBM2 memory, HBM2 IP (controller, PHY, custom I/O), ASIC core blocks and total power consumed by the 2.5D HBM2 ASIC SiP at various voltages and temperatures.

In conclusion, Open-Silicon's silicon proven HBM2 IP subsystem in TSMC's 16nm FF+ is ready for SoCs enabling next generation high bandwidth applications. Customers can minimize the integration risk by using an HBM2 IP subsystem (controller + PHY + I/O) from a single vendor. In addition, learnings from the silicon validation and interoperability with HBM2 memory die stack will be incorporated in Open-Silicon's next generation AXI (Advanced eXtensible Interface) based HBM2 IP subsystem development targeting 2.4Gbps per pin data rates, and beyond, in TSMC's 7nm FF technology.

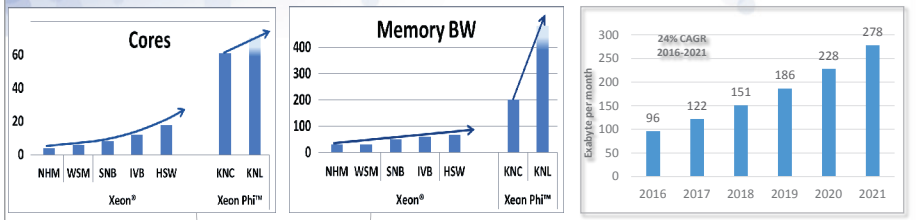


High Bandwidth Memory (HBM2) IP Subsystem Silicon Validation And Interoperability With HBM2 Memory Die Stack

Vinay Somanache
Principal Architect, Design
Co-authors Naveen HN and Sachin Ghadi

Increase in Cores, BW and Data: Driving New Silicon Markets

HPC And Networking Applications



HPC
Source : Intel

Networking
Source : Cisco

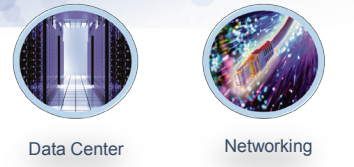
Custom Processors

ASICs

Specialized Memory

- ***High-Bandwidth Memory***

HBM Applications & Comparison


Data Center

Networking


Artificial Intelligence

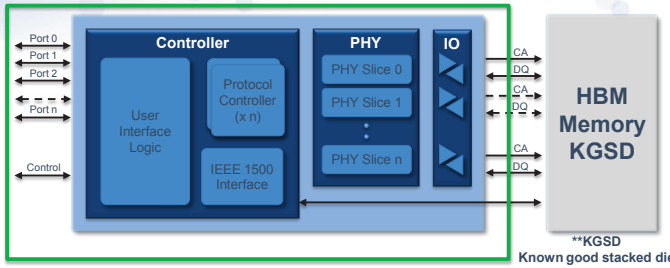

AR and VR


Cloud Computing


Neural Networks

	DDR4	GDDR5	HBM2
Bus Width (per chip)	16-bit	32-bit	1024-bit
Clock Speed	3.2Gbps	7Gbps	2.0Gbps
Bandwidth (per chip)	6.4Gbytes	28Gbytes	256Gbytes
Power Efficiency (Gbits per sec/pin)	1mW	-	0.33mW
Voltage	1.2V	1.35V, 1.5V	1.2V
Package Type	Discrete	Discrete	Si-Interposer

HBM2 IP Subsystem Features (16nm)



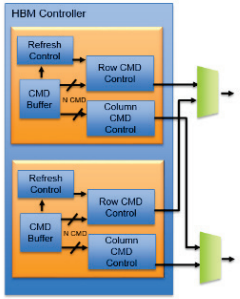
- Supports full HBM subsystem (HBM Ctrl + HBM PHY + HBM IOs)
- Supports up to 2Gbps/pin data rate
- Supports up to 8 Channels (16 Pseudo Channels)
- Supports up to 256GBytes of Total Bandwidth
- Compliant to JEDEC HBM 2.2 Specification
- Supports up to 5mm interposer trace length
- Supports AXI and native host interface
- Supports multiple in-built test & diagnostic features

High Performance HBM Controller Architecture

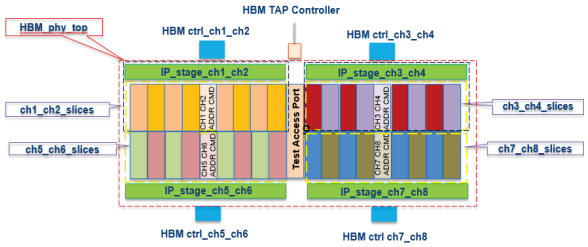
Conventional DDR shortcomings
- Shared row and column command address bus
- Active/pre-charge or read/write commands can only be executed one at a time

Semi-independent row and column command interface
- Allows simultaneous active pre-charge and read/write commands

Controller architecture takes the advantages of
- Isolating the row and column operations
- The row state machine prioritizes the banks to be opened for read/write operations
- Effectively it can open multiple banks required in a pool of commands
- The column state machine prioritizes between read/write commands



HBM PHY & I/O

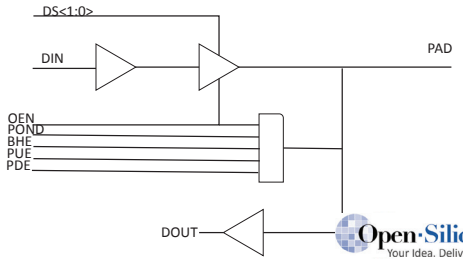


PHY Hierarchy

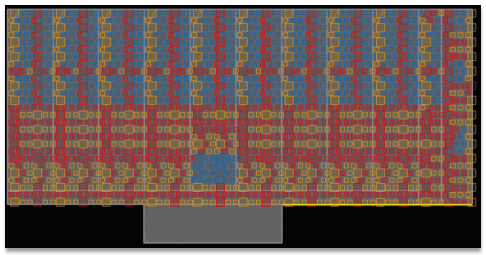
Scalable and optimized PHY

Custom die-to-die 2.5D I/O

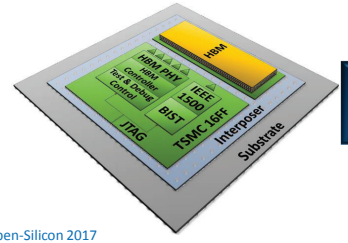
CMOS IO driver, 1GHz/2Gbps DDR with light output loading



Probing Scheme for 2.5D HBM2 ASIC SiP



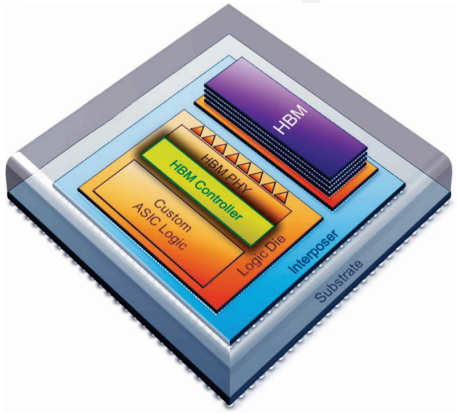
- Signal uBump
- Gnd uBump
- Power uBump
- Gnd Probe-Pad
- Power Probe-Pad



Testability feature deployed through probing mechanism for debug of ASIC SiPs with HBM memory embedded on a 2.5D interposer die

2.5D HBM2 ASIC SiP For Silicon Validation Parameters

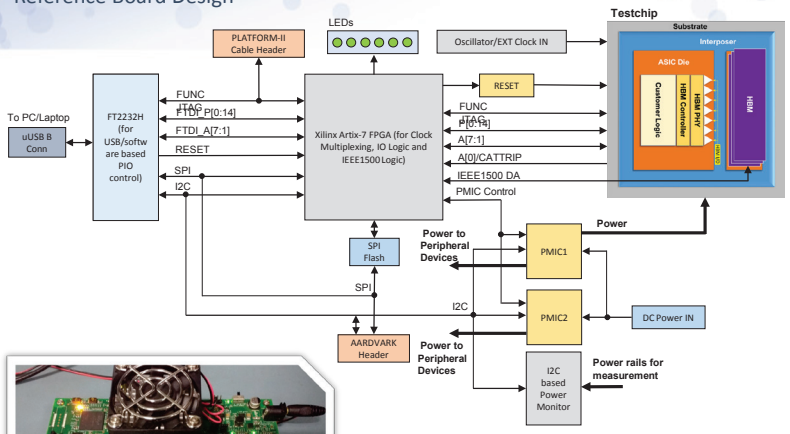
- ASIC Technology **TSMC 16 nm**
- Interposer Technology : **TSMC 65 nm**
- SiP process : **TSMC CoWoS (2.5D)**
- **HBM2 Memory**
- **Open-Silicon HBM2 IP Subsystem**
- Core Voltage: **0.8V**
- IO Voltage : **1.2V**



This Platform Demonstrates
Open-Silicon HBM2 IP Subsystem ⇔ HBM2 Die-Stack Interoperability

2.5D HBM2 ASIC SiP Validation/Evaluation Platform

Reference Board Design



Open·Silicon Your Idea. Delivered.™

2.5D HBM2 ASIC SiP Validation/Evaluation Platform

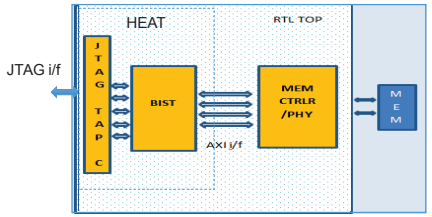
Hardware Enable Algorithm Testing (HEAT)

HEAT is used for On-chip Silicon Validation of Memory subsystems

- Mem Protocol Controller
- Mem PHY
- IOs

HEAT Features

- JTAG interface for programming on chip
- At speed Functional Testing (1Ghz) of the Memory subsystem
  - At speed Traffic Generation(mimicking system traffic)
- Data integrity check
- Performance Measurement
  - Min and Max latency with transaction details
  - Average latency
  - Timeout status with configurable timeout value

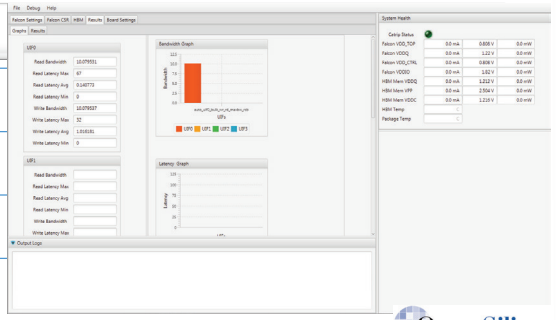


Open·Silicon Your Idea. Delivered.™

2.5D HBM2 ASIC SiP Validation/Evaluation Platform

Software Interface To Validate HBM Platform

- Jtag2Custom IP converts JTAG transactions into custom interfaces
- Direct Control of CSR Registers in the HBM controller IP and Test Chip Registers
- Displays system health ,voltage ,temperature ,power of memory and controller
- Custom interface supports
  - Built-in data generator engine to drive data traffic
  - Can generate either a fixed or variable data pattern using a pseudo random binary sequence (PRBS)
  - Executes random/increment/decrement/walking zero/walking one addressing patterns
  - Compares and gathers Bandwidth latency, data mismatch results



Open·Silicon Your Idea. Delivered.™

2.5D HBM2 ASIC SiP Silicon Validation Methodology

Bandwidth and Latency

- Validation platform will capture test status statistics, bandwidth and latency

Signal Integrity

- Calibration report on a valid data window validating signal integrity

SiP Characterization Across Voltage & Temperature

- Data rates across voltage and temperate range

Open·Silicon Your Idea. Delivered.™

Performance: Bandwidth and Latency

Simulation Results at 2.0 Ghz			*16nm
	Average Latency	**Bandwidth**	**Data Utilization**
Incremental Reads (PS Ch)	32	14.25 GBps	89.0625
Incremental Writes (PS Ch)	50	14.37 GBps	89.8125
True Random Reads (PS Ch)	189	1.45 GBps	9.0625
True Random Writes (PS Ch)	200	1.45 GBps	9.0625
Silicon Results at 2.0 Ghz			*16nm
	Average Latency	**Bandwidth**	**Data Utilization**
Incremental Reads (PS Ch)	30	14.15 GBps	88.4375
Incremental Writes (PS Ch)	44	14.15 GBps	88.4375
True Random Reads (PS Ch)	176	1.44 GBps	9
True Random Writes (PS Ch)	197	1.45 GBps	9.0625

*Due to AC parameters of the memory die and interface protocol timing, there are non-data cycles that results in Data utilization drop
*Random Read/Writes means true random in all bits of address which results in significant increase in average latency and data utilization drop. It is recommended to have deterministic data patterns result in better data utilization. Experiment of randomness was done to get worst case data bandwidth

SiP Characterization Across Voltage & Temperature
2.5D HBM2 ASIC SiP

Data rate @ 1.2V

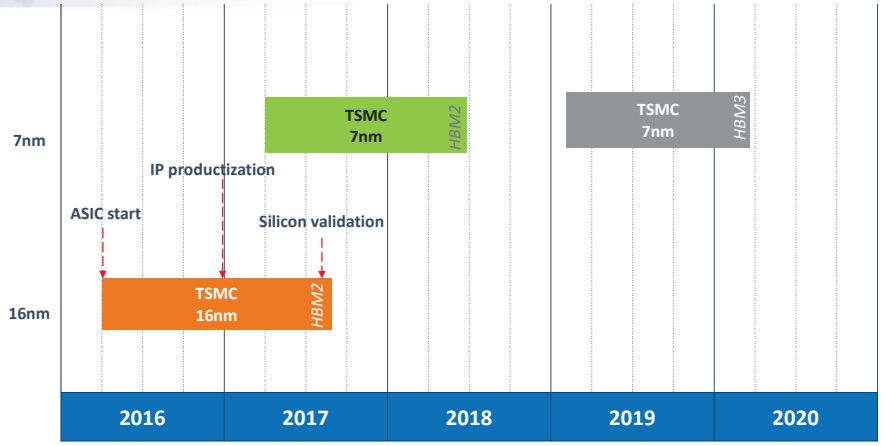
Voltage/Temp	-10°C	25°C	85°C
1.14V	1.6Gbps	1.6Gbps	1.6Gbps
1.2V	1.6Gbps	1.6Gbps	1.6Gbps
1.26V	1.6Gbps	1.6Gbps	1.6Gbps

Data rate @ 1.35V

Voltage/Temp	-10°C	25°C	85°C
1.28V	2.0Gbps	2.0Gbps	2.0Gbps
1.35V	2.0Gbps	2.0Gbps	2.0Gbps
1.42V	2.0Gbps	2.0Gbps	2.0Gbps

Required data rates are achieved across voltage and temperature range as per specification

HBM IP Subsystem Roadmap
TSMC Foundry

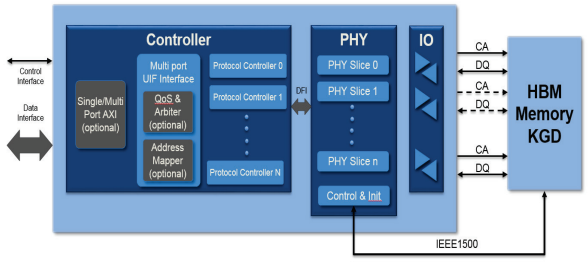


Next Generation HBM2 IP subsystem (7nm)

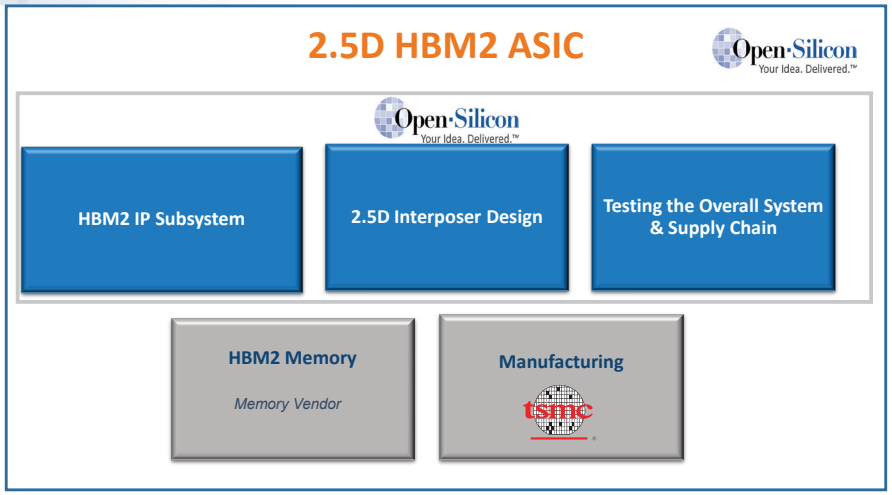
AXI (Advanced eXtensible Interface) based HBM2 IP subsystem development

Targeting 2.4Gbps per pin data rates, and beyond, in TSMC' s 7nm FF technology

- Supports up to 2.4Gbps/pin data rate
- Supports up to 8 Channels (16 Pseudo Channels)
- Supports up to >300GBytes of Total Bandwidth
- Supports full DFI compliant Controller and PHY interface
- Supports multi-port AXI interface
- Supports different schemes of arbitration and scheduling (QoS)
- Supports different address mapping modes



HBM2 ASIC SiP Solution



Summary

Open-Silicon's silicon proven HBM2 IP subsystem in TSMC's 16nm is ready for SoCs enabling next generation high bandwidth applications.

Customers can minimize the integration risk by using an HBM2 IP subsystem (controller + PHY + I/O) from a single vendor

Silicon validation demonstrates functional validation and interoperability between Open-Silicon's HBM2 IP subsystem and HBM2 memory

Data rates of 1.6Gbps/2Gbps demonstrated on HBM2 SiP solution in TSMC 16nm supporting up to 5mm interposer trace length

Open-Silicon's next generation multi port AXI (Advanced eXtensible Interface) based HBM2 IP subsystem development targeting 2.4Gbps per pin data rates, and beyond, in TSMC's 7nm technology

THANK YOU